图书情报工作
第64卷 第17期 2020年9月

多源信息融合的微博查询似然模型*

■ 吴树芳[1] 张雄涛[2] 朱杰[3]
[1] 河北大学管理学院 保定 071000 [2]北京科技大学东凌经济管理学院 北京 100083
[3] 中央司法警官学院信息管理系 保定 071000

摘 要：[目的/意义]查询似然模型存在零概率问题，融合多源信息对模型进行扩展，不仅可以解决零概率问题，还可以实现对全局信息的差异化处理，降低噪声。[方法/过程]通过LDA主题挖掘和历史微博兴趣挖掘，分别获取初始微博的主题相关信息和兴趣相关信息，并将二者与全局信息融合，用于改进初始微博的语言模型估计，从而得到扩展的微博查询似然模型。运用网络爬虫工具从新浪微博爬取数据，并通过实证研究验证扩展模型的有效性。[结果/结论]实验结果表明：与已有的查询似然模型扩展方法相比，新模型具有较好的检索性能。

关键词：多源信息 微博检索 查询似然模型 主题信息 作者兴趣
分类号：G254
DOI：10.13266/j.issn.0252-3116.2020.17.012

1 引言

随着移动互联网的进一步发展，微博逐步成为人们生产、分享和消费信息的重要平台。为解决海量微博导致的信息过载问题，微博检索已成为用户获取有效信息的重要途径。微博检索不同于传统的文本检索，其检索对象具有碎片化、网络术语化、符号化等特点，其排序原则不仅要考虑查询和微博的相似度，还要考虑其他信息（如兴趣信息、时间信息等），故直接将传统文本检索模型用于微博检索不妥。查询似然模型是当前主流的微博检索模型，其相似度计算包括文档先验概率和文档语言模型估计（即词项在文档语言模型中的概率分布），其中文档语言模型估计是否准确直接影响到模型的检索性能，为解决数据稀疏性导致该估计值可能出现的零概率问题[1]，相关学者对查询似然模型进行了系列扩展研究，扩展内容主要围绕文档语言模型的估计展开。

在传统文本检索领域，文档语言模型估计的扩展研究大致分为两个阶段：①通过引入全局信息对语言模型进行估计，如Jelinek-Merrer（JM）方法[2]和Dirichlet Prior（DIR）方法[3]。此类方法虽有效解决了传统语言模型估计中的零概率问题，但由于未对全局信息进行差异化处理，导致大量噪声信息的引入，从而影响了信息检索的准确性；②通过融合全局信息和其他相关信息对语言模型进行估计，如X. Liu等[4]利用聚类信息对全局信息进行修订，提出一种融合聚类信息和全局信息的语言模型估计方法，实现对传统查询似然模型的扩展。T. Tao等[5]将内容近邻信息与全局信息进行融入，得到一种改进的查询似然模型。实证研究结果表明：相较于第一阶段直接引入全局信息的扩展方法，第二阶段的方法可以更准确地对语言模型进行估计，从而有效地实现对查询似然模型的扩展。

上述研究主要针对的是传统文本检索，考虑到微博和传统文本的不同，相关研究者结合微博的特点对微博查询似然模型展开了扩展研究，基本思路为：首先结合微博的特点确定微博相关信息，然后融合相关信息和全局信息改进微博查询似然模型。例如：M. Efron等[6]考虑到微博较短，将其视为查询，并结合相关反馈方法得到微博的相关信息，通过融合微博的相关信息和全局信息对微博语言模型进行估计，得到改进的微博查询似然模型；李锐等[7]考虑到微博的时效性和交互性，基于用户的历史微博和交互信息获取相关微博，将获取的相关微博和全局微博融合，得到一种改进的微博语言模型估计方法；M. Efron等[8]利用

* 本文系国家社会科学基金项目"网络信息治理视域下社交网络不可信用户识别研究"（项目编号：17BTQ068）研究成果之一。
作者简介：吴树芳（ORCID：0000-0001-6587-812X），教授，博士，博士生导师；张雄涛（ORCID：0000-0002-2134-9602），博士研究生，通讯作者，E-mail：zhangxiongtao1@163.com；朱杰（ORCID：0000-0002-5698-135X），副教授，博士。
收稿日期：2019-12-16 修回日期：2020-05-04 本文起止页码：114-122 本文责任编辑：杜杏叶

Hashtag 获取相关微博,提出一个融合 Hashtag 和全局信息的微博查询似然模型。Hashtag 是微博中一个特殊标签,具有相同 Hashtag 的微博属于同一个话题,利用 Hashtag 包含的信息可有效获得当前微博的相关微博。

综上,微博语言模型估计是微博查询似然模型中的关键项,估计准确与否直接影响微博检索的性能,而如何获取有效的相关微博信息是实现准确估计的关键。基于此,论文在已有研究的基础上,综合考虑微博自身信息、全局信息、主题信息以及作者兴趣信息 4 个方面,多维度获取相关微博,提出一种多源信息融合的微博查询似然模型。在信息检索领域,主题挖掘是获取文本语义信息的重要手段[9],LDA(Latent Dirichlet Allocation)主题模型[10]自 2003 年被提出之后,已被广泛应用于主题挖掘,论文将依据 LDA 模型进行主题挖掘,获取微博的主题相关信息。此外,用户兴趣挖掘是实现个性化检索的关键技术之一[11],有效的兴趣挖掘方法是获得微博相关信息的另一种途径。

2 研究设计

多源信息的获取是论文所提出扩展模型的关键,其中自身信息即为微博本身,全局信息为当前可以使用的所有信息,二者易于获取。对于主题相关信息,论文采用实证研究的方法,基于 LDA 模型进行主题挖掘,通过计算基于主题分布的距离,获得主题相关微博集(基于主题的语义信息)。对于兴趣相关信息,首先基于用户的历史微博集挖掘其兴趣,然后通过兴趣相关度计算获得兴趣相关微博集。上述信息中,全局信息是当前可用的微博全集,该类信息可以有效解决数据稀疏性问题,但其缺点是会引入很多噪声。降低这些噪声的方法就是提高全局信息中更能体现初始微博内容相关信息的权重,这样噪声权重排序就会后移,从而降低噪声引入的概率,主题相关信息和兴趣相关信息的引入可以达到上述降噪的目的。

本文的研究框架如图 1 所示,主要工作包括以下 3 个方面:

(1)微博 d_i 主题相关信息的获取:基于 LDA 主题模型进行主题挖掘,将微博文本表示为 m 个主题下的概率分布向量,并基于微博文本的主题分布差异计算微博相关度,获得微博 d_i 基于主题信息的相关微博集 T。

(2)微博 d_i 兴趣相关信息的获取:依据微博 d_i 所属作者的历史微博挖掘作者兴趣,为体现兴趣的动态性,给出了兴趣词的动态权重计算方法。通过计算每条微博和作者兴趣的相似度,获得作者的兴趣微博集 I。

(3)多源信息融合:将 T、I 和微博全集融合,平滑初始微博 d_i,重新估计词项在微博中的概率分布,得到扩展的微博查询似然模型。

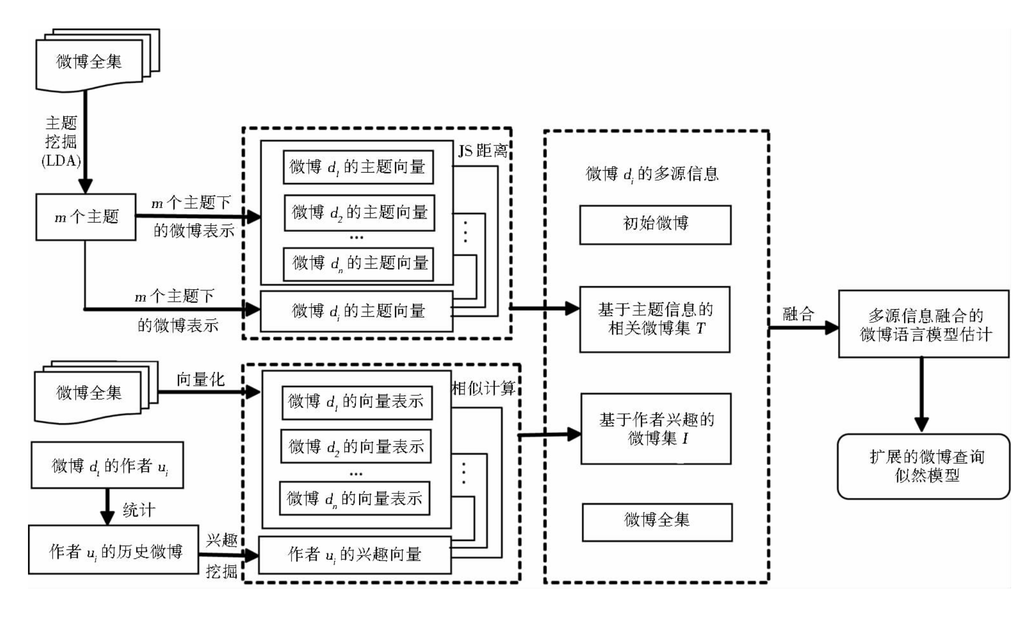


图 1 微博查询似然模型扩展研究框架

3 传统查询似然模型

在信息检索领域中,语言模型将文本表示成词语的联合概率分布。J. M. Ponte 和 W. B. Croft[12]提出的查询似然模型是语言模型应用于信息检索的经典模型,计算公式如下:

$$P(d_i|q)=\log P(d_i)+\sum_{k\in V}c(k,q)\log P(k|M_{d_i})$$ 公式(1)

其中,q 表示查询,d_i 表示文档,k 表示词语,V 表示所有词语的集合,M_{d_i}表示文档语言模型;$P(d_i|q)$表示在查询 q 的条件下检索到文档 d_i 的概率;$P(d_i)$表示文档 d_i 的先验概率,一般采用等概率度量,即假定待检索文档的先验概率是相等的;$c(k,q)$表示词语 k 在查询 q 中出现的次数;$p(k|M_{d_i})$为词语 k 在文档语言模型 M_{d_i} 中的概率分布,即文档语言模型的估计,计算方法如公式(2)所示:

$$P(k|M_{d_i})=P_{ml}(k|M_{d_i})=\frac{c(k,d_i)}{|d_i|}$$ 公式(2)

公式(2)中,$P_{ml}(k|M_{d_i})$表示采用极大似然估计法计算 $P(k|M_{d_i})$,$c(k,d_i)$,表示词语 k 在文档 d_i 中出现的次数,$|d_i|$表示文档 d_i 中包含词语的个数。k 表示词项全集中的词,如果文档 d_i 中不包括该词,则会出现零概率问题,导致 $\log P(\cdot)$计算无意义,且文档越短零概率问题越严重。实际情况中,词语 k 虽然不在文档 d_i 中出现,但是其相关词如果在 d_i 中出现,该概率值不应该为 0,故解决零概率问题的关键是找到相关信息,有效平滑初始文档 d_i,提出改进的概率估计方法计算 $p(k|M_{d_i})$,这也是本文的研究出发点。

4 多源信息融合的微博查询似然模型

为克服传统查询似然模型的不足,准确估计词项在文档中的概率分布,提高基于查询似然模型的微博检索的综合性能,本文在已有研究的基础上,采用主题相关信息(属于语义相关)和兴趣相关信息(属于个性化信息)对全局信息进行差异化处理,平滑初始微博,提出一个扩展的微博查询似然模型。

4.1 基于主题信息的相关微博集获取

LDA 模型是当前主流的文本主题挖掘方法,通过 LDA 模型对文本训练后,文本可从词语空间映射到主题空间,实现文本的语义表示[13]。本文采用 LDA 主题模型对微博文本进行建模:首先利用 Python 中的 Gensim 工具包训练得到 m 个主题,然后将每条微博表示为在 m 个主题下的概率分布,获得如表 1 所示的微博 – 主题概率分布矩阵。其中,n 表示微博全集中微博的条数,m 表示微博主题的个数(其值由实验获取),d_i 表示第 i 条微博,$Topic_j$ 表示第 j 个主题,P_{ij}表示第 i 条微博在第 j 个主题上的分布概率。

表 1 微博 – 主题概率分布矩阵

微博 \ 主题	$Topic_1$	…	$Topic_j$	…	$Topic_m$
d_1	P_{11}	…	P_{1j}	…	P_{1m}
…	…	…	…	…	…
d_i	P_{i1}	…	P_{ij}	…	P_{im}
…	…	…	…	…	…
d_n	P_{n1}	…	P_{nk}	…	P_{nm}

经过上述训练后,微博 d_i 可被表示为由不同主题下的概率分布组成的主题向量,即:

$$d_i=(P_{i1},P_{i2},\cdots,P_{im})$$ 公式(3)

在上述表示的基础上,我们采用 JS 距离[14]计算任意两条微博 d_i 和 d_j 的主题相关度,计算公式如下:

$$JS(d_i,d_j)=\frac{1}{2}\left[KL(d_i,\frac{d_i+d_j}{2})+KL(d_j,\frac{d_i+d_j}{2})\right]$$ 公式(4)

公式(4)中,$KL(\cdot)$用于度量两个量之间的非对称距离,计算方法如公式(5)所示,$\frac{d_i+d_j}{2}$表示微博 d_i 和 d_j 在 m 个主题上的分布均值。JS 距离越大,则微博之间的分布差异越大,相关度越小。本文依据该值,将 JS 距离由小到大排序,选取 $Top-N_1$(N_1 的取值通过实验获得)个微博组成与当前微博主题相关的微博集合 T。

$$KL(d_i,d_j)=\sum_{r=1}^{m}P_{ir}\log\frac{P_{ir}}{P_{jr}}$$ 公式(5)

公式(5)中,P_{ir}表示微博 d_i 在主题 $Topic_r$ 上的概率分布,P_{jr}表示微博 d_j 在主题 $Topic_r$ 上的概率分布。

4.2 基于作者兴趣的相关微博集获取

历史微博可有效体现用户兴趣[15],本文依据历史微博挖掘作者兴趣,计算每条微博和作者兴趣的相似度,最终通过阈值判断获得作者的兴趣微博,所有兴趣微博组成基于作者兴趣的相关微博集 I。假设用户 u_i 是微博 d_i 的作者,该作者的历史微博集为 D,k_j 为历史微博集中的任意一个词语,则 k_j 的初始权重计算公式为:

$$w_{k_j-original}=\frac{n_{ji}}{|d_i|}\times\log\frac{|D|}{|\{r:k_j\in d_r\}|}$$ 公式(6)

公式(6)中,$w_{k_j-original}$表示词语 k_j 的初始权重,n_{ji}表示词语 k_j 在微博 d_i 中出现的次数,$|d_i|$表示微博 d_i 中包含的词语个数,$|D|$表示历史微博集中的微博条数,$|\{r:k_j\in d_r\}|$表示历史微博集中包含词语 k_j 的微博数。

考虑到微博用户的兴趣会随时间而逐渐衰减,本

吴树芳，张雄涛，朱杰. 多源信息融合的微博查询似然模型[J]. 图书情报工作,2020,64(17):114－122.

文基于词语所属微博的发布时间对词语的权重进行更新。依据指数衰减的思想,更新后词语 的权重计算公式为:

$$w_{k_j-new} = w_{k_j-original} \times e^{-\mu V_t} \quad \text{公式(7)}$$

其中,w_{k_j-new} 表示词语更新后的权重,Δt 表示微博发布时间与历史微博集中最新时间的距离,μ 为指数衰减参数,本文依据 J. Choi 的实验结果将其设置为 0.02。

由于词语 k_j 在不同微博中的权重可能不同,本文将 k_j 在整个历史微博集中权重的平均值作为该词语在历史微博集中的权重,即:

$$w_{k_j-D} = \frac{\sum_{|D|} w_{k_j-new}}{|D|} \quad \text{公式(8)}$$

其中,w_{k_j-D} 表示历史微博集 D 中词语 k_j 的权重,$|D|$ 表示作者历史微博集中微博的条数。通过上述计算,可以得到作者历史微博集中每个词语的权重。本文依据词语权重选取 $Top-N_2$ 个词语表示作者兴趣(N_2 的取值在实验部分说明),例如作者 u_i 的兴趣可表示为:

$$u_{i-interest} = \{k_1, k_2, \cdots, k_{N_2}\} \quad \text{公式(9)}$$

获得用户的兴趣表示后,采用公式(10)计算微博集中任意一条微博 d_r 和作者兴趣 $u_{i-interest}$ 的相似度,如下式:

$$sim(d_r, u_{i-interest}) = \frac{\sum_{N_2}(w_{k_j} - D \times |d_{r-k_j}|)}{|d_r|} \quad \text{公式(10)}$$

其中,$sim(d_r, u_{i-interest})$ 表示微博 d_r 和作者兴趣 $u_{i-interest}$ 的相似度,N_2 表示作者兴趣表示词的个数,w_{k_j-D} 为依据公式(8)计算的作者兴趣词 k_j 的权重,$|d_{r-k_j}|$ 表示作者兴趣词 k_j 在微博 d_r 中出现的次数,$|d_r|$ 表示微博 d_r 包含的词语个数。选取相似度大于阈值 δ 的微博作为作者的兴趣微博,构成作者 u_i 的兴趣微博集 I。

4.3 扩展的微博查询似然模型

经上述处理后,可获得微博 d_i 的基于主题信息的相关微博集 T 和微博 d_i 所属作者的兴趣微博集 I,通过融合词语 k 在原微博 d_i 中的分布 M_{d_i}、在主题相关微博集 T 中的分布 M_T、在兴趣相关微博集 I 中的分布 M_I 以及在全局信息中的分布 M_C,得到如公式(11)、(12)所示的微博查询似然模型:

$$P(d_i|q) = \log P(d_i) + \sum_{k \in V} c(k,q) \log P(k|M_{d_i})_{improve} \quad \text{公式(11)}$$

$$P(k|M_{d_i})_{improve} = \beta_1 P_{ml}(k|M_{d_i}) + \beta_2 P_{ml}(k|T_T) + \beta_3 (P_{ml}(k|T_I) + \beta_4 P_{ml}(k|M_C) \quad \text{公式(12)}$$

观察公式(11)可以发现,论文提出的扩展模型主要改进了文档语言模型估计 $P(k|M_{d_i})$ 的计算,避免了传统查询似然模型存在的不足。其中,$P(k|M_{d_i})_{improve}$ 表示改进后的微博语言模型估计,M_T 表示与微博 d_i 主题相关微博集 T 构建的语言模型,M_I 表示依据微博 d_i 所属作者的兴趣微博集 I 构建的语言模型,$P_{ml}(k|M_{d_i}$ 表示语言模型 M_{d_i} 的最大似然估计,$P_{ml}(k|M_T)$ 表示语言模型 M_T 的最大似然估计,$P_{ml}(k+M_I)$ 表示语言模型 M_I 的最大似然估计,$P_{ml}(k|M_C)$ 表示语言模型 M_C 的最大似然估计,以上估计均采用公式(2)计算。$\beta_i(i=1,2,3,4)$ 为调和参数,且 $\sum_{i=1}^{4}\beta_i=1$。全局信息($P_{ml}(k|M_C)$)的加入可避免零概率的问题,因为词语 k 可能不属于某微博 d_i,但肯定来源于全局。融合全局信息的弊端是噪声的引入,为解决该问题,论文依据主题相关信息($P_{ml}(k|M_T)$)和兴趣相关信息($P_{ml}(k|M_I)$)对全局信息进行差异化处理,提高相关词的概率,在特征选择时,小概率词将会被去掉,从而有效地避免了噪声的引入。

本文采用层次分析法[17]确定公式(12)中的平滑参数 $\beta_i(i=1,2,3,4)$。首先依据 1－9 重要程度判断表,对各项的重要程度进行两两比较,得到如表 2 所示的判定矩阵;然后基于判断矩阵,计算得到最大特征根为 4.138 9,特征向量为(0.54,0.25,0.15,0.06),一致性指标为 0.046 3,一致性比例为 0.051 4。由于一致性比例小于 0.1,判断矩阵通过一致性检验,本文将 β_i 的取值分别确定为:0.54、0.25、0.15、0.06。

表 2 判定矩阵

平滑参数	β_1	β_2	β_3	β_4
β_1	1	3	4	6
β_2	1/3	1	2	5
β_3	1/4	1/2	1	4
β_4	1/6	1/5	1/4	1

5 实证研究

5.1 实验数据

新浪微博是我国当前最具权威性的微博平台,本文采用网络爬虫工具爬取 661 845 条新浪微博数据,依据查询似然模型构建了微博检索系统。爬取的数据包括微博文本内容、微博发布时间和微博作者三类信息。为避免无效数据的干扰,本文参照 TREC 会议(Text Retrieval Conference)[18]的评测要求并结合本文的实验需要,对爬取的新浪微博数据进行如下处理:①去除已失效或只含有表情符号的微博;②去除长度小于 30 个字符的微博;③将数据集中的所有繁体字转换为简体

字;④采用 Python 中 jieba 包对每条微博进行分词处理,并结合哈工大整理的《停用词表》对分词处理后的微博文本进行去停用词处理。

对爬取的微博语料进行上述处理后,本文参照信息检索领域中小型测试集的构建方法[19],选取 17 010 条微博作为微博检索系统的文档集,构建了 5 个查询及相关查询集,每个查询的相关性采用 Pooling 方法[20]进行标注。查询及其相关文档数、不相关文档数如表 3 所示:

表 3 微博检索测试集

查询	内容	相关文档数	不相关文档数
查询 1	快乐大本营二十周年生日	192	16 818
查询 2	演员的诞生章子怡战队	236	16 774
查询 3	时尚芭莎明星慈善基金晚会	225	16 785
查询 4	电视剧烈火如歌的大结局	85	16 925
查询 5	妈妈是超人中的黄圣依	131	16 879

5.2 评价标准

本文采用前 k 个返回结果的准确率($P@k$)和平均倒数排名(MRR)对微博检索性能进行评价。其中,$P@k$ 是 TREC 微博检索任务中官方指定的评价指标,MRR[21]是近年来比较流行的评价指标,其在准确率的基础上考虑了位置因素,可有效衡量相关文档的位置信息。的计算公式如下:

$$P@k = \frac{1}{k}\sum_{j=1}^{k} r_j \qquad \text{公式(13)}$$

$$MRR = \frac{1}{|R|}\sum_{i=1}^{|R|}\frac{1}{rank_i} \qquad \text{公式(14)}$$

公式(13)中,k 表示前 k 个检索结果,如果检索结果的第 j 篇文档是相关的,则 $r_j=1$,否则 $r_j=0$。本文取值 $k=30$,因为微博网页检索的前两页共包含 30 个检索结果。公式(14)中,$|R|$ 为相关文档的总数,$rank_i$ 为返回结果中第 i 个相关文档的位置,MRR 的值越高,则相关文档在结果列表中越靠前,检索性能越好。模型解释如表 4 所示:

表 4 模型简写及其解释

不同查询似然模型简写	模型解释
LM	基于微博本身的微博查询似然模型
LM-JM	基于微博本身、全集微博的查询似然模型
LM-JM-Topic	基于微博本身、全集微博、主题相关微博的查询似然模型
LM-JM-Interest	基于微博本身、全集微博、作者兴趣微博的查询似然模型
LM-JM-Topic-Interest	基于微博本身、全集微博、作者兴趣微博、主题相关微博的查询似然模型

5.3 实验及分析

本文实验分两部分:第一部分为相关参数的设定;第二部分为微博检索采用不同查询似然模型时对应的检索性能对比。其中,第二部分实验中涉及的微博查询似然模型的简写及解释见表 4。

5.3.1 相关参数分析

本文实验涉及的相关参数包括:微博主题个数 m,主题相关微博个数 N_1,作者兴趣表示词个数 N_2,作者兴趣度阈值 δ。这些参数均通过反复实验进行确定。

(1)微博主题个数 m 的确定。本文通过计算微博文本的困惑度(perplexity)[22]来确定微博主题的较优个数,困惑度越小,表示模型生成文本的能力越强,性能越好,其计算公式为:

$$perplexity = \exp\left(\frac{-\sum_{i=1}^{M}\log P(k_{d_i})}{\sum_{i=1}^{M} N_{d_i}}\right) \qquad \text{公式(15)}$$

其中,M 表示全集微博 D 中包含的微博数目,d_i 为 D 中的任意一条微博,N_{d_i} 表示微博 d_i 中包含的词语数目,k_{d_i} 表示微博 d_i 中的词语,$P(k_{d_i})$ 表示微博 d_i 中词语 k_{d_i} 出现的概率,可采用公式(16)计算:

$$P(k_{d_i}) = \sum_z P(z)\cdot P(k_{d_i}|z) \qquad \text{公式(16)}$$

其中,z 表示微博 d_i 涉及的某个主题,由于微博文本属于典型的短文本,涉及的主题个数有限,依据本文爬取的实验数据规模,将分别计算主题个数为 1,2,3,4,5,6,7,8,9,10 时的模型困惑度,得到如图 2 所示的不同主题个数下模型的困惑度。由图 2 可以看出,当主题个数 $m=4$ 时,LDA 模型的困惑度较小,故本文将主题个数设定为 4。

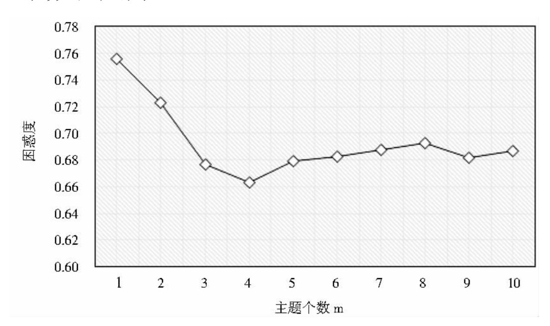


图 2 不同主题个数下 LDA 模型的困惑度

(2)基于主题信息的相关微博集中微博个数 N_1 的确定。采用适量的主题相关微博集进行平滑可有效提高微博语言模型估计的准确性,数量太少平滑效果不明显,数量太多可能会引入噪声。为获得合理的基于主题信息的相关微博数 N_1,本文首先分别以数据量间隔为 10、100 和 1 000 为单位,进行实证研究,结果发现:当间隔为 10 时,主题信息无法充分利用;当间隔为

1 000 时,引入大量噪声。因此,本文将微博数量间隔设置为 100。实验基于 LM-JM-Topic 方法对微博语言模型进行估计,在 N_1 分别取 100、200、300、400、500、600、700、800、900、1 000 时,分别计算 5 个查询在微博检索系统中的 P@30,实验结果如表 5 所示:

表 5　不同主题相关微博数下 5 个查询的 P@30

P@30	100	200	300	400	500	600	700	800	900	1 000
查询 1	0.433	0.467	0.567	0.533	0.533	0.533	0.500	0.467	0.433	0.433
查询 2	0.533	0.600	0.567	0.567	0.533	0.533	0.567	0.567	0.533	0.533
查询 3	0.400	0.567	0.633	0.633	0.633	0.467	0.433	0.400	0.433	0.433
查询 4	0.367	0.367	0.367	0.333	0.300	0.300	0.267	0.267	0.267	0.233
查询 5	0.500	0.533	0.567	0.567	0.533	0.533	0.433	0.400	0.400	0.367
平均值	0.447	0.507	**0.540**	0.527	0.506	0.473	0.440	0.420	0.413	0.400

从表 5 中可以看出,当 N_1 =300 时,5 个查询的 P@30 平均值达到较高值,这说明选取 Top-300 条主题相关微博可以较为有效地对微博语言模型进行估计,这里的 300 属于粗略较优取值,如果想获得更为准确的数据可以采用上述实验方法,将间隔值调小,重复上述实验。值的注意的是,查询 2 中基于主题信息的相关微博增加到 700,查询 3 中基于主题信息的相关微博增加到 900 时,P@30 的值出现小幅反复。通过分析这些微博发现:对应数量的微博与原始微博虽然主题相关度较低,但在作者交互、发布时间分布等方面具有较高相关性,故综合相关度较高,导致出现上述小幅反复现象,但是,这些反复并未超过 P@30 最大值,且从平均 P@30 值来看,小幅反复在整体上并不影响 P@30 的递减趋势,故本文将 N_1 初步设定为 300。

(3)作者兴趣表示词个数 N_2 的确定。本文采用公式(6)计算作者历史微博集中的词语权重,并对其进行排序,然后选取 Top-N_2 个词语表示作者兴趣。N_2 过小,作者兴趣难以被充分表示,N_2 过大,对象表示的区分度较低。为选取适量的词语表示作者兴趣,本文采用 AUS(平均用户满意度)指标[14]确定作者兴趣表示词的数量。具体过程为:首先随机选取微博数据集中 10 个微博作者,并采用专家小组法对 10 个作者的兴趣词语进行标注,选择出可以表示每个作者兴趣的兴趣词集(规模小于或等于 40);然后分别计算这 10 个作者的 Top-N_2 的 AUS 值,并选取较高的 AUS 对应的兴趣词个数作为本文所需确定的 N_2。AUS 的计算如公式为:

$$AUS = \frac{\sum_{m=1}^{M} \frac{n_{u_m}}{N_2}}{M} \qquad 公式(17)$$

公式(17)中,M 表示随机选取的作者数(本文取 M=10),n_{u_m} 表示由专家标注的作者 u_m 的兴趣词个数,N_2 表示从兴趣词列表中选取的词语个数。本文分别令 N_2 取值 10、20、30、40、50、60、70、80、90、100 得到图 3 所示的不同兴趣词个数与 AUS 之间的关系。从图 3 中可以看到,当作者的兴趣表示词个数为 40 时,AUS 值较高,故本文初步设定 N_2 =40。

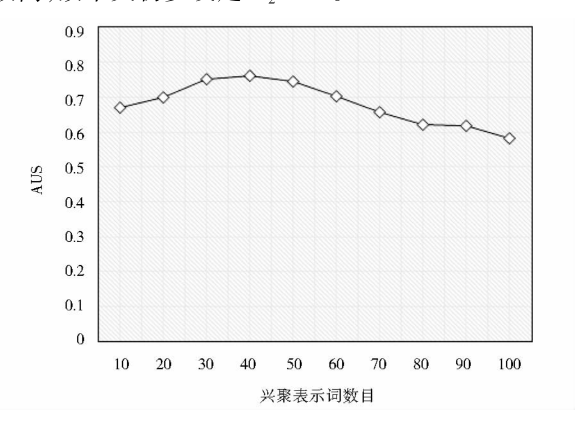


图 3　不同兴趣表示词个数对应的 AUS

(4)作者兴趣度阈值 δ 的确定。为获得合理的作者兴趣度阈值,实验基于 LM-JM-Interest(该方法主要考虑了兴趣相关信息)查询似然模型,在 δ 分别取 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9 时,分别计算 5 个查询在微博检索系统中的 P@30,实验结果见表 6。

从表 6 中可以看出,当 δ=0.8 时,5 个查询的 P@30 平均值达到较高值。值的注意的是,查询 2 在兴趣度阈值降低到 0.4 时,查询 5 的兴趣度阈值降低到 0.2 时,P@30 的值出现小幅反复。这些小幅反复并未超过 P@30 最大值,且从平均值来看,其在整体上并不影响 P@30 的递减趋势,故本文将 δ 初步设定为 0.8。

5.3.2　LM-JM-Topic, LM-JM-Interest 和 LM-JM-Topic-Interest 性能比较

本文依据 P@30 和 MRR 两个指标对 LM-JM-Topic、LM-JM-Interest 和 LM-JM-Topic-Interest 3 种查询似然模型进行性能比较。

表 6 不同兴趣微博数下 5 个查询的 *P*@30

P@30	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
查询 1	0.467	0.467	0.467	0.467	0.467	0.433	0.400	0.367	0.367
查询 2	0.567	0.600	0.567	0.533	0.567	0.600	0.567	0.533	0.500
查询 3	0.400	0.467	0.433	0.367	0.333	0.333	0.300	0.267	0.233
查询 4	0.300	0.333	0.333	0.300	0.300	0.300	0.267	0.233	0.233
查询 5	0.633	0.767	0.767	0.733	0.733	0.700	0.667	0.700	0.567
平均值	0.473	0.527	0.513	0.480	0.480	0.473	0.440	0.420	0.380

（1）*P*@30 比较。在微博检索系统中，分别采用 LM-JM-Topic、LM-JM-Interest 和 LM-JM-Topic-Interest 作为表示模型时，模型检索结果对应的 *P*@30 如图 4 所示：

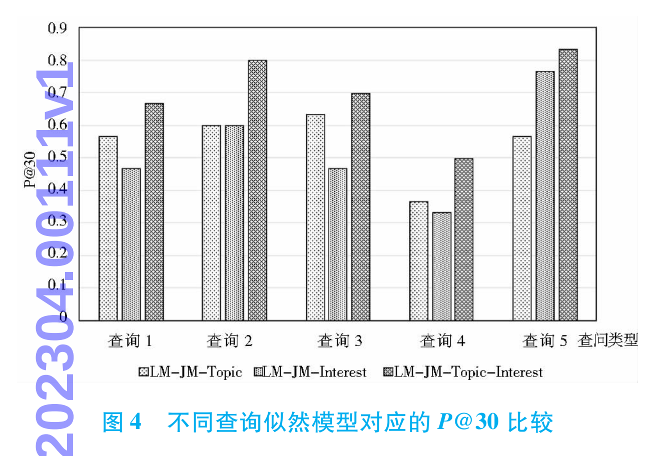


图 4 不同查询似然模型对应的 *P*@30 比较

从图 4 可以发现：采用 LM-JM-Topic-Interest 查询似然模型时，5 个查询在微博检索系统中对应的 *P*@30 值均高于其他两种方法，这说明本文最终提出的 LM-JM-Topic-Interest 查询似然模型相比于 LM-JM-Topic 查询似然模型和 LM-JM-Interest 查询似然模型，可以得到更为准确的估计值，进而提高微博检索系统的查准率。

（2）*MRR* 比较。针对测试集合中的 5 个查询，LM-JM-Topic、LM-JM-Interest 和 LM-JM-Topic-Interest 3 种模型对应的文档检索排名指标 *MRR* 值见图 5。从图 5 可以看出，采用 LM-JM-Topic-Interest 查询似然模型对文档进行检索时，5 个查询在微博检索系统中对应的 *MRR* 值均高于其他两种方法，这说明本文最终提出的 LM-JM-Topic-Interest 模型相比于 LM-JM-Topic 和 LM-JM-Interest 可以使相关文档的排名更靠前。

综上，采用 LM-JM-Topic-Interest 查询似然模型对微博进行检索时，5 个查询在微博检索系统中的 *P*@30 指标和 *MRR* 值均优于其他两个模型。产生这种结果的原因是：LM-JM-Topic-Interest 方法既考虑了主题相关微博的信息，又考虑了作者的兴趣信息，而其他两种方法引入的相关信息相对片面，进而导致微博语言模型估计的准确性不足，影响微博检索系统的综合性能。

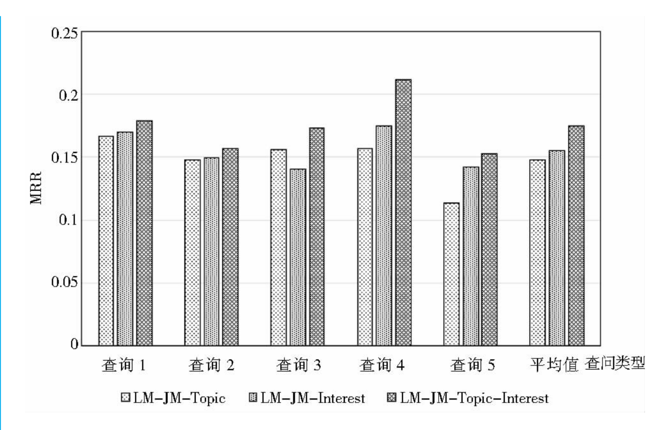


图 5 不同查询似然模型对应的 *MRR* 比较

5.3.3 LM，LM-JM 和 LM-JM-Topic-Interest 3 种方法比较

此部分实验用于比较本文最终提出的 LM-JM-Topic-Interest 查询似然模型和传统查询似然模型 LM、基于全局信息扩展的查询似然模型 LM-JM 的性能。

（1）*P*@30 比较。图 6 为采用 3 种模型对测试集合中的 5 个查询进行检索时，得到的 *P*@30 比较图。从图 6 可以发现：采用 LM-JM-Topic-Interest 模型对微博进行检索时，5 个查询在微博检索系统中对应的 *P*@30 值均高于其他两种方法，这说明本文最终提出的 LM-JM-Topic-Interest 方法相比于 LM 方法和 LM-JM 方法，可以使微博检索系统得到更高的查准率。

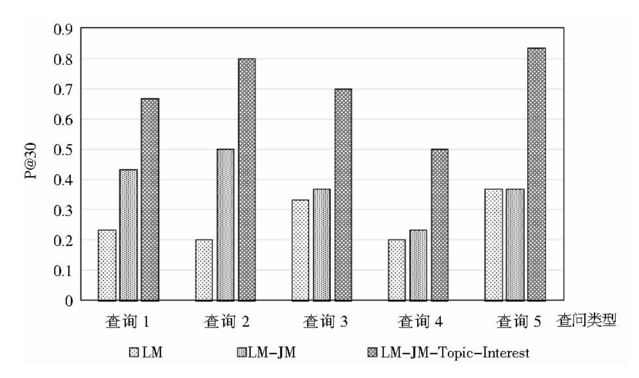


图 6 查询在不同模型中的 *P*@30 比较

（2）*MRR* 值比较。图7为3种模型的 *MRR* 值。观察图7可以发现,采用 LM-JM-Topic-Interest 模型对微博进行检索时,5个查询在微博检索系统中对应的 均高于其他两种方法,这说明本文最终提出的 LM-JM-Topic-Interest 方法相比于 LM 方法和 LM-JM 方法可以使相关文档的排名更靠前。

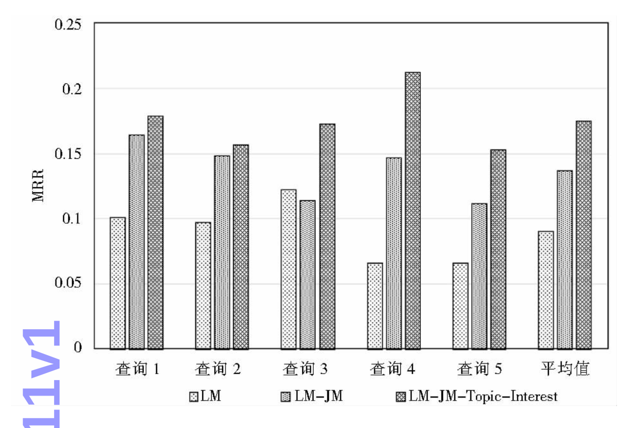


图7　查询在不同模型中的 *MRR* 比较

综合以上:采用 LM-JM-Topic-Interest 模型进行微博检索时,5个查询在微博检索系统中均能获得比其他两种模型更好的检索性能。因此,LM-JM-Topic-Interest 模型优于其他两种模型。产生这种结果的原因是:传统查询似然模型仅考虑了微博本身的信息,存在由于数据稀疏性导致的零概率问题。基于全局信息的平滑方法虽然解决了零概率问题,但过多的补充信息会引入噪声数据。而本文提出的微博查询似然模型利用主题相关信息和作者兴趣信息对全局信息进行了差异化处理,可以有效提高相关词的概率,降低噪声词的概率,进而提高微博语言模型估计的准确性,提高了微博检索的性能。

6　结语

考虑到已有查询似然模型存在的不足,本文综合利用微博自身信息、主题信息、作者兴趣信息以及全集微博信息,提出了一种多源信息融合的微博查询似然模型。与已有研究相比,本研究虽在一定程度上提高了微博检索的性能,但尚存不足之处,未来研究拟围绕以下内容展开深入研究:①本文工作针对的是离线形式的微博数据,而实际微博数据是以数据流的形式实时更新,故未来研究中我们拟结合在线学习思想对微博查询似然模型进行改进。②本文主要结合了4个方面的信息改进查询似然模型,但有利于微博语言模型估计的信息还有其他多个方面(如:时间信息,作者交互信息等),故未来研究将深入挖掘其他有效信息,进一步提高微博语言模型估计的准确性,进而提高微博检索的性能。

参考文献:

[1] 吴树芳，张雄涛，朱杰. 融合用户兴趣和混合估计的微博检索模型[J]. 情报学报，2019，38(4)：411-419.

[2] KATZ S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE transactions on acoustics speech & signal processing, 2003, 35(3): 400-401.

[3] GANGULY D, ROY D, MITRA M, et al. A word embedding based generalized language model for information retrieval[C]//Proceedings of the 38th international ACM SIGIR conference. Santiago: ACM, 2015:795-798.

[4] LIU X, CROFT W B. Cluster-based retrieval using language models[C]//Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. Sheffield: ACM, 2004:186-193.

[5] TAO T, WANG X, MEI Q, et al. Language model information retrieval with document expansion[C]//Proceedings of the human language technology conference of the north American chapter of the ACL. New York: Association for Computational Linguistics, 2006: 407-414.

[6] EFRON M, ORGANISCIAK P, FENLON K. Improving retrieval of short texts through document expansion[C]//Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. Portland: ACM. 2012.

[7] 卫冰洁，史亮，王斌. 一种融合聚类和时间信息的微博排序新方法[J]. 中文信息学报，2015，29(3)：177-183,189.

[8] EFRON M. Hashtag retrieval in a microblogging environment [C]//Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. Geneva: ACM, 2010:787-788.

[9] 张小鹏，吕学强，李卓，等. LDA 与词汇链相结合的主题短语抽取方法[J]. 小型微型计算机系统，2018，39(11):107-113.

[10] BLEI D M, NG A Y, JORDAN M I, LAFFERTY J. Latent dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993-1022.

[11] JIANG Y, XU Y, SHAO L A personalized microblog search model considering user-author relationship[C]//Proceedings of international conference on data science in cyberspace. Changsha: IEEE, 2016:508-513.

[12] PONTE J M, CROFT W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 1998:275-281.

[13] 刘德喜，付淇，韦亚雄，等. 基于多重增强图和主题分析的社交短文本检索方法[J]. 中文信息学报，2018，32(3)：110-

119.

[14] 唐晓波，房小可. 一种面向微博的查询扩展方法［J］. 图书情报工作，2014，58(1)：130－135.

[15] 熊才伟，曹亚男. 基于发文内容的微博用户兴趣挖掘方法研究［J］. 计算机应用研究，2018(6)：63－71.

[16] CHOI J，CROFT W B. Temporal models for microblogs［C］//Proceedings of the 21st ACM international conference on Information and knowledge management. Maui：ACM，2012：2491－2494.

[17] VAIDYA O S，KUMAR S. Analytic hierarchy process：An overview of applications［J］. European journal of operational research，2006，169(1)：1－29.

[18] LIN J，ROEGIEST A，TAN L，et al. Overview of the TREC 2016 real-time summarization track［C］//Proceedings of the 25th text retrieval conference. Maryland：TREC，2016.

[19] 徐建民，王平. 小型中文信息检索测试集的构建与分析［J］. 情报杂志，2009，28(1)：13－16.

[20] CORMACK G V，PALMER C R，CLARKE L A. Efficient construction of large test collections［C］//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne：ACM，1998：282－289.

[21] WANG Y，HUANG H，FENG C. Query expansion based on feedback concept model for microblog retrieval［C］// Proceedings of the 26th international conference on World Wide Web. Perth：International world wide web conferences steering committee，2017：559－568.

[22] 关鹏，王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究［J］. 现代图书情报技术，2016(9)：42－50.

作者贡献说明：

吴树芳：论文思路的提出、撰写及修改，实证指导；
张雄涛：论文的撰写，实证研究；
朱杰：实验数据的收集、处理及论文修改。

Microblog Query Likelihood Model Based on Multi-Source Information Fusion

Wu Shufang[1] Zhang Xiongtao[2] Zhu Jie[3]
[1] School of Management，Hebei University，Baoding，071000
[2] Dongling School of Economics and Management，University of Science and Technology，Beijing 100083
[3] Department of Information Management，the Central Institute for Correctional Police，Baoding 071000

Abstract：［**Purpose/significance**］Due to the existence of zero probability problem in the query likelihood model，we propose to extend the model by multi-source information fusion，which not only solves zero probability problem，but also achieves the differential processing of global information to reduce the introduction of noise.［**Method/process**］Topic related information and interest related information were obtained based on LDA topic mining and historical Microblog interest mining respectively，then we integrated them with global information to modify the evaluation of the original Microblog's language model. Finally，an extended microblog query likelihood model is obtained. We used the web crawler tools to crawl data from Sina Weibo to verify the effectiveness of the extended model by empirical study.［**Result/conclusion**］Experimental results indicate that our model can achieve better retrieval performance.

Keywords：multi-source information　microblog retrieval　query likelihood model　topic information　author interest